

Thesis for the Master of Science

A Modification–Aware Framework for Fast Open  
Modification Spectral Library Search

*Younghee Seo*

Graduate School of Hanyang University

*February 2025*

Thesis for Master of Science

A Modification–Aware Framework for Fast Open  
Modification Spectral Library Search

Thesis Supervisor: Prof. Eunok Paek

A Thesis submitted to the graduate school of  
Hanyang University in partial fulfillment of the requirements  
for the degree of Master of Science

*Younghee Seo*

*February 2025*

Department of Artificial Intelligence  
Graduate School of Hanyang University

This thesis, written by Younghee Seo,  
has been approved as a thesis for the Master of Science.

*February 2025*

Committee Chairman: Heejin Park

(Signature) 

Committee member: Eunok Paek

(Signature) 

Committee member: Seungjin Na

(Signature) 

Graduate School of Hanyang University

# Table of Contents

Abstract	iii
1 Introduction	1
2 Background	4
3 Method: Modification-Aware Candidate Selection	9
3.1 Complementary Spectrum Generation	9
3.2 Four Spectral Database Searches	9
4 Data	10
5 Results	12
5.1 Overall PSM Identification	12
5.2 Analysis by Modification Position	15
5.3 Correlation Analysis of Shifted Dot Product and Vector Similarity	18
5.4 trade off between search space and search time.	19
5.5 PTM type	19
5.6 Performance Comparison on HEK293 Raw Data	21
6 Conclusion	22
Reference	23
Abstract in Korean	25

## Figures

Figure 1	Overview of ANN-SoLo	3
Figure 2	Creating complementary spectrum and pipeline for our method	7
Figure 3	Performance comparison between ANN-SoLo and our method	11
Figure 4	Psm analysis and examples of spectrum that are found by our method	13
Figure 5	Correlation between inner product and shifted dot product	17
Table 1	PSM comparison on HEK293raw data	20
Table 2	correct sequence comparison on HEK293 raw data	21

# **Abstract**

## **A Modification-Aware Framework for Fast Open Modification Spectral Library Search**

**Younghee Seo**

**Department of Artificial Intelligence**

**The Graduate School**

**Hanyang University**

The analysis of peptides using mass spectrometry is a fundamental yet complex task in proteomics. Recently, various algorithms have been developed to interpret mass spectrometry data more effectively. Among these, spectral library search through Open Modification Search (OMS) has gained prominence, particularly with the introduction of ANN-SoLo, which enhances the efficiency of OMS by clustering candidate peptides using an Approximate Nearest Neighbor (ANN)-based algorithm. However, ANN-SoLo does not incorporate modification information during clustering, potentially overlooking modified peptides. To address this limitation, we propose a novel approach that integrates complementary spectra to account for modification information during clustering. By ensuring that both modified and unmodified spectrum pairs are grouped within the same cluster, we improved the identification of modified peptides. Furthermore, our modification-aware method refined the selection of candidate peptides, allowing for more accurate identification based on the location of

potential modification sites. This approach not only accelerated the search process but also improved the accuracy of OMS in spectral library searches, offering significant benefits to proteomic research.

## 1. Introduction

The analysis of peptides using tandem mass spectrometry (MS/MS) is a fundamental yet complex task in proteomics, essential for identifying and characterizing peptides within biological samples. Recently, various algorithms have been developed to enhance the interpretation of MS/MS data. Among these, spectral library search has gained prominence for its efficiency and sensitivity. By comparing query spectra directly with previously validated spectra stored in spectral libraries, this method not only improves computational efficiency but also increases sensitivity compared to traditional sequence database searches.

A major challenge in proteomics is the detection of post-translational modifications (PTMs), as conventional methods are limited in their ability to identify modifications unless they are pre-specified within the database. Standard approaches typically require a user to specify a limited set of expected modifications, restricting the search space and potentially overlooking unexpected or novel PTMs. To address these limitations, Open Modification Search (OMS) has become a valuable tool for PTM analysis, as it allows for the identification of peptides with a broad range of modifications without prior specification [1,2].

OMS works by applying a wide precursor mass window, often set as large as 500 m/z, which enables the matching of modified peptides with unmodified counterparts based on precursor mass differences. However, while OMS broadens the search space for PTM discovery, this approach increases computational demands due to the large volume of candidate spectra that must be processed.



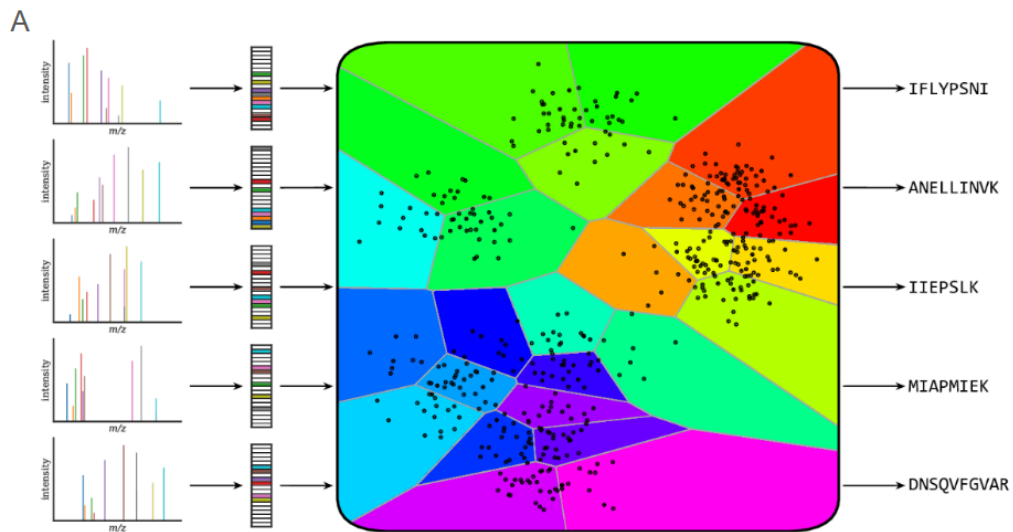
To enhance OMS efficiency, ANN-SoLo applied an Approximate Nearest Neighbor (ANN) algorithm to cluster and prioritize candidate spectra [3], significantly reducing the number of comparisons required. Through feature hashing [4], ANN-SoLo converted each spectrum's  $m/z$  and intensity values into vectors—hereafter referred to as "hash vectors"—allowing for efficient candidate selection in high-dimensional search spaces. Although ANN-SoLo improved the speed of OMS, it did not consider modification information when generating clusters, limiting its ability to accurately identify modified peptides.

To address this limitation, we generated hash vectors using a complementary spectrum approach that integrates modification information into the clustering process [5]. This complementary spectrum was created by using each fragment ion  $m/z$  and the precursor mass to calculate the theoretical position of its complementary ion (e.g., given a b-ion, its corresponding y-ion position is inferred, and vice versa), allowing fragment ions including modified residues to align with the positions of unmodified counterparts. To improve candidate selection further, we conducted searches across four separate spectral databases: a complementary spectrum considering modification locations, an intensity-weighted complementary spectrum, the original spectrum, and an intensity-weighted version of the original spectrum. This multi-database candidate selection strategy enabled more accurate identification of modified peptides while preserving computational efficiency.

To evaluate the performance of modified spectra searches, we needed a ground truth dataset of modified spectra queries, which we obtained using MODa [6]. MODa, a tag-based database search tool designed to identify post-translationally modified peptides, provided peptide-spectrum matches (PSMs) that served as ground truth data for our evaluation. Our method not only identified a greater number of PSMs compared to ANN-SoLo but also retrieved additional PSMs

beyond the ground truth dataset, demonstrating improved identification of modified peptides. Furthermore, this enhanced performance was achieved without compromising runtime efficiency.

We analyzed differences based on modification positions, focusing on how modifications at different regions of the peptide sequence influence identification of modified peptides. Our analysis demonstrated that our method was particularly effective at identifying peptides with modifications located near the C-terminal, where ANN-SoLo often struggles due to its inability to account for the dominance of y-ions and the extensive shifts caused by such modifications. This highlights the strength of our method in handling near C-terminal modifications, offering a more robust approach to these challenging cases.



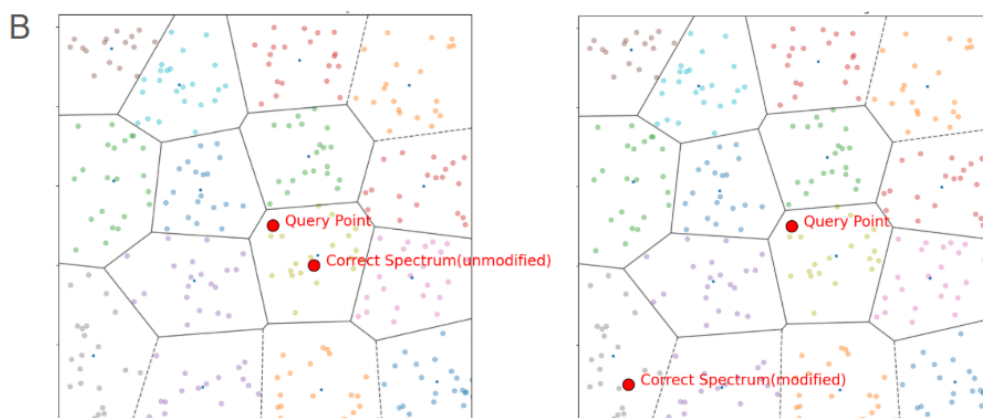


Figure1. A) This figure is from the ANN-SoLo paper [3]. B) This figure illustrates two scenarios within the ANN-SoLo algorithm: the left side represents an ideal situation, while the right side depicts a case where a modified spectrum is a query, and the corresponding unmodified spectrum vector lies in a cluster far from the query's cluster. This situation highlights a limitation of ANN-SoLo, as it does not consider modifications during vectorization. Our method is designed to overcome this limitation by addressing such cases effectively.

## 2. Background

### 1. ANN-SoLo

ANN-SoLo is a tool optimized for fast and accurate open modification spectral library searching, designed to address the extensive search space introduced using a wide precursor mass window. To enhance search efficiency, ANN-SoLo employs a two-level search. Initially, it performs a standard search to identify

unmodified peptides or peptides with modifications already present within the database. For any remaining query spectra not matched in this initial phase, ANN-SoLo proceeds to an open search, expanding the search space to identify potential modifications(Figure 1A).

ANN-SoLo speeds up the open search process by using Approximate Nearest Neighbor (ANN) indexing, specifically leveraging the Faiss library for efficient clustering and candidate selection. The tool first converts each spectrum into a hashed vector, capturing the essential m/z and intensity information while simplifying the dimensionality for faster comparisons.

In ANN-SoLo, the Faiss library is used to cluster spectra based on similarity [7], with each cluster represented by a centroid. During candidate selection, the centroids are initially compared with the query spectrum to limit the search space, thus reducing computational demands. After determining the closest centroids, ANN-SoLo retrieves candidates from the corresponding clusters for further matching. Two key parameters control this clustering and retrieval process:

- $N_{list}$ : Specifies the number of centroids. In ANN-SoLo, this is set to 256, providing a balance between processing speed and candidate accuracy.
- $N_{probe}$ : Determines how many clusters are probed when retrieving candidates. The default in ANN-SoLo is 128, which ensures that a sufficient number of clusters are considered for high recall without excessive computation.

After selecting candidates, Faiss library provides candidate indexes that are similar to the query spectrum vector. ANN-SoLo uses a shifted dot product between the given library spectrum and query spectrum. These spectra are not in a hashed vector form, but real spectra are obtained by candidate index. By using

this ANN-based strategy, ANN-SoLo efficiently narrows down the search space, but it does not account for modifications during candidate selection. This limitation can lead to missing modified spectra that should ideally be included as candidates (Figure 1B). According to Faiss recommendations, optimal values for large-scale libraries are  $N_{list} \approx \sqrt{database\ size}$  and  $N_{probe} \approx N_{list}/4$ . For the MassIVE-KB spectral library used in ANN-SoLo, a more suitable setting would be an  $N_{list}$  of 4096, which can further improve both speed and accuracy.

After the open search process, the FDR control step is performed. ANN-SoLo uses group FDR for this purpose [8]. In group FDR, PSMs with the same mass difference are grouped together, and FDR is calculated within each group. The minimum group size is set to 20, and if a group contains fewer than 20 PSMs, those with ungrouped mass differences are combined into a single group, and FDR is calculated for this combined set. Using a 1% FDR with this FDR control introduces certain problems. Specifically, when dealing with fewer than 100 PSMs, even a single decoy can result in exceeding the 1% FDR threshold. However, since this study focuses solely on modifying the candidate selection process for comparison, the same FDR control was applied. Nonetheless, further algorithmic improvements are required to address these issues. As this study focuses solely on modifying the candidate selection process, the FDR control mechanism was not altered. Instead, the default parameters of ANN-SoLo were used for FDR control.

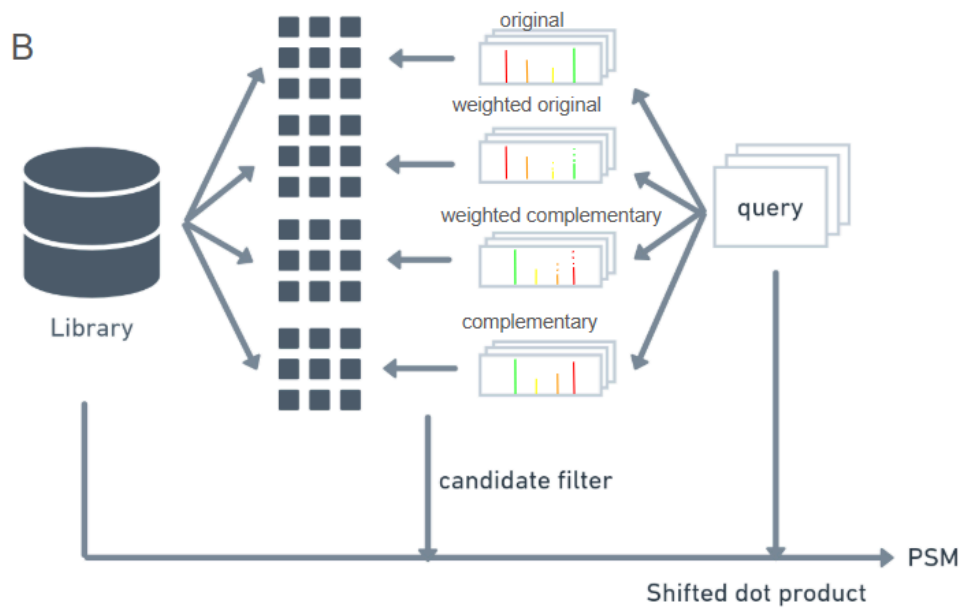
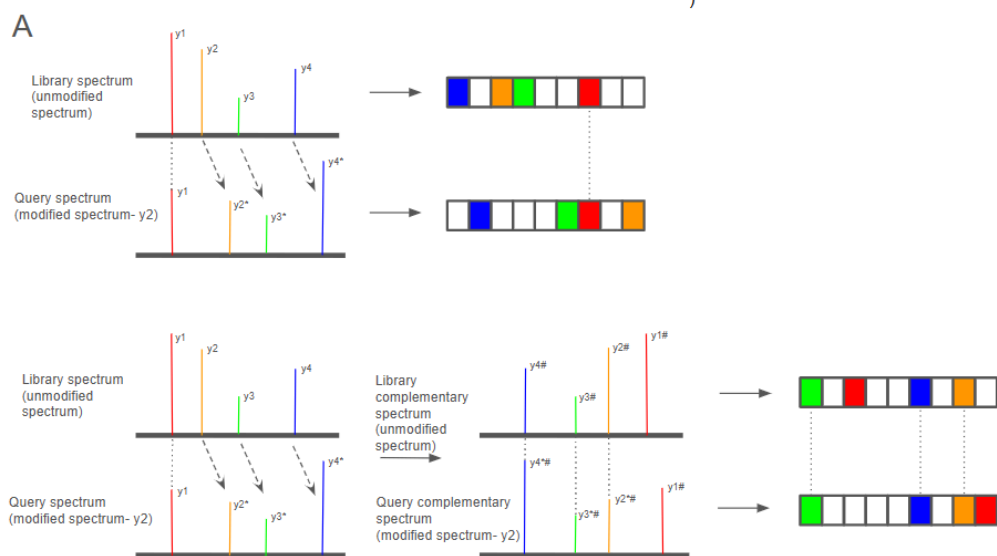


Figure 2.A) Comparison of Spectrum Vector Generation in ANN-SoLo and the Proposed Method Using Complementary Spectrum for Modification-Aware Matching: top) This figure illustrates how ANN-SoLo generates spectrum vectors using feature hashing when creating candidates. When the query contains a modification, indicated by arrows in the figure, the similarity between the two generated spectrum vectors—the query spectrum vector and the library spectrum vector—tends to be low. As a result, it is less likely that such a spectrum will be selected as a candidate. bottom). This figure represents the method we proposed. Before generating the spectrum vector, a complementary spectrum was first created. By converting this complementary spectrum into a spectrum vector, we ensure that the peaks affected by the modification can still be compared properly. As a result, the query spectrum vector and the library spectrum vector are better matched. B) Pipeline for our method. First, we vectorize all spectra in the library using four different approaches, as described in method 2, and store them in separate vector databases. These databases are clustered and indexed using Faiss. When searching a query spectrum, the query spectrum is vectorized using the same four approaches applied to the library spectra. Once the candidate selection is completed, the candidate spectrum indices are filtered using an open search mass tolerance of 500 Da to extract the final candidate spectra from the library. The shifted dot product is then used to calculate the final PSMs, ensuring accurate identification.

### **3. Method**

#### **Modification-Aware Candidate Selection**

##### **3.1 Complementary Spectrum Generation**

To overcome ANN-SoLo's limitation of ignoring modification information, we developed a complementary spectrum approach that aligns modified peaks to their corresponding unmodified positions. Using each peak and the precursor mass, we calculate its complementary ion position (e.g., treating a peak as a b-ion allows us to infer its y-ion position, and vice versa). This adjustment aligns the modified spectrum more closely with the unmodified library spectrum, improving the probability of accurate matching(Figure 2A).

##### **3.2 Four Spectral Database Searches**

First, we created four vector databases from the spectral library using Faiss, each designed to consider the modification site when selecting candidates:

1. The first database stores vectors created from the original spectrum.
2. The second database applies a penalty to peaks at high  $m/z$  values in the original spectrum by reducing their intensity by half.
3. The third database, like the second, applies the same intensity reduction to peaks at high  $m/z$  values in the complementary spectrum.
4. The fourth database stores vectors created from the complementary spectrum.

Next, the query spectrum is also converted into four vectors using the same method as for the databases. Each of these query vectors is searched against the corresponding database. If the query has no modification or the modification



site is located near the peptide N-terminus, its best match will likely come from the first database. If the modification site is located around the middle of the peptide, its best match will probably come from the second or third database. If the modification site is near the C-terminus, its best match will likely come from the fourth database.

Finally, the candidate spectra identified through this process are scored using the shifted dot product method [9], as in the original ANN-SoLo, to obtain the final peptide-spectrum matches (Figure 2). This approach works because, during peptide fragmentation in MS/MS data, y-ions are observed more frequently than b-ions. When analyzing the library data, we examined the coverage of y-ions and b-ions. The results showed that the total intensity of y-ions (the sum of the intensities of y-ion peaks) was 3.6 times greater than that of b-ions.

## **4. Data**

The query spectra utilized in this study comprise 225,336 modified spectra derived from the HEK293 cell line, identified using MODa. The peptide sequences assigned by MODa are considered as the ground truth for these spectra. For the spectral library, we used 2,140,865 high-resolution human peptide spectra obtained from label-free HCD experiments, sourced from the MassIVE-KB database[10].

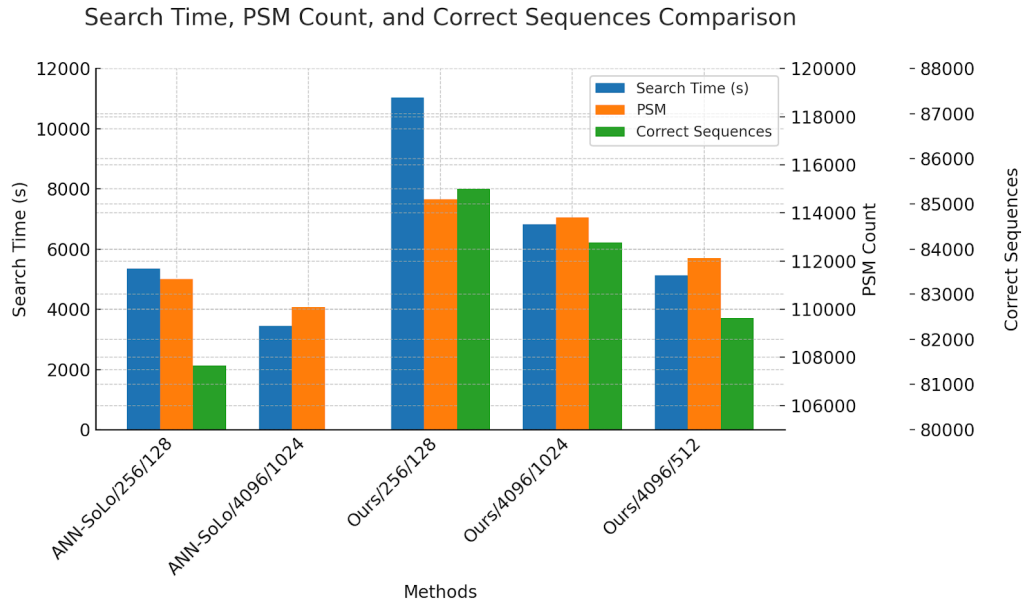


Figure3. This figure illustrates the performance differences between our method and the existing ANN-SoLo. The x-axis represents the method / # of lists (Faiss param.) / # of probes (Faiss param.). Our method not only identifies more PSMs but also demonstrates that the PSMs found align with the ground truth data. The dataset used for this figure is the results of an open search conducted on 225,336 query spectra after excluding the 83,441 PSMs obtained during the standard search (the first stage of the cascade search, which identifies unmodified peptides or modifications present in the database). As a result, the open search was performed on the remaining 141,895 query spectra.

## 5. Results

### 5.1 Overall PSM Identification

To evaluate the performance of our method compared to ANN-SoLo, we analyzed the peptide-spectrum matches (PSMs) identified by each approach. Specifically, we focused on the total number of PSMs and the fraction of PSMs containing the correct peptide sequences. The parameters for ANN-SoLo were set to their default values of  $N_{list}=256$  and  $N_{probe}=128$ , whereas our method utilized  $N_{list}=4096$  and  $N_{probe}=512$ , as recommended by the original report of Faiss. ANN-SoLo identified 194,694 PSMs, while our method identified 195,559 PSMs. Among the identified PSMs, ANN-SoLo found 148,767 correct sequences, whereas our method identified 149,821 correct sequences. This demonstrates that our method identified more correct sequences than the increase in total PSMs, highlighting its meaningful improvement(Figure 3).

Additionally, our method achieved these results without compromising search time, maintaining a comparable performance while delivering more accurate outcomes.



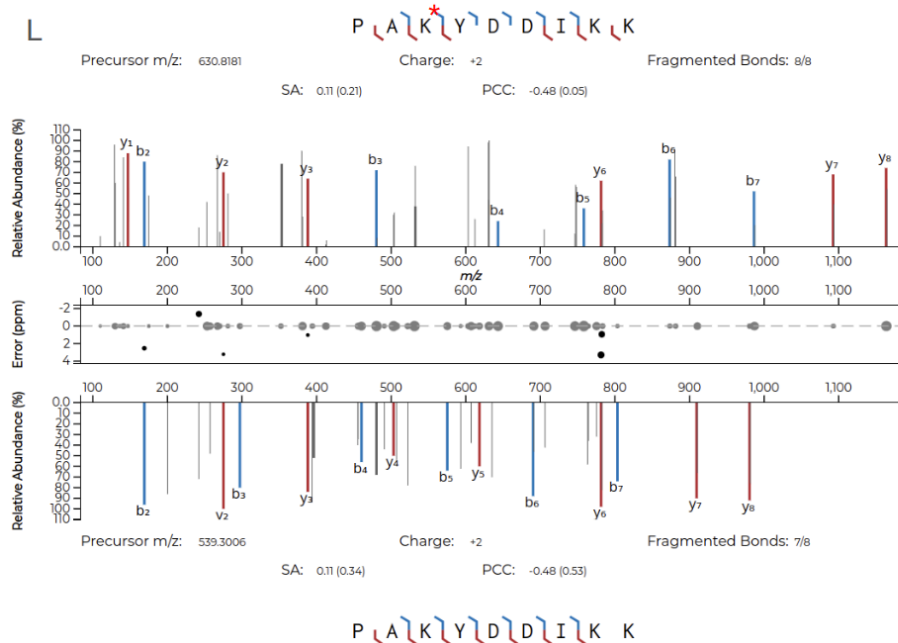
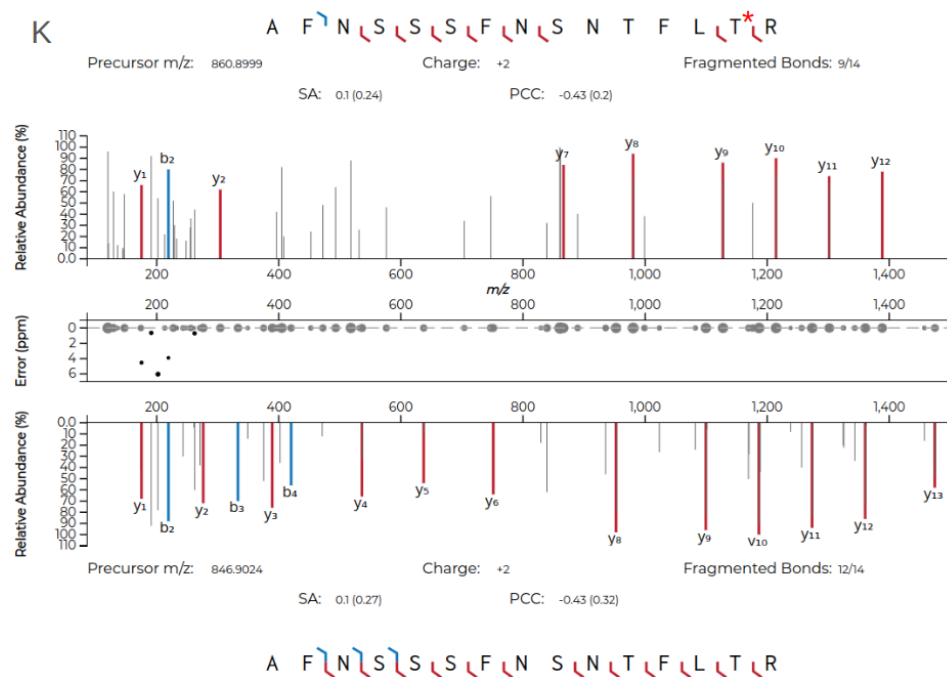


Figure 4. The diagrams labeled **a~e** show the results of comparing the PSMs obtained by ANN-SoLo and our method. Diagram **a** illustrates the total PSMs identified from the 225,336 query spectra, while **b** shows the PSMs for spectra with two or more modifications. Diagrams **c~e** present cases with a single modification, categorized based on its position: near the N-terminal (**c**), in the middle region (**d**), and near the C-terminal (**e**). Similarly, **f~j** show the results based on correct sequences identified (ground truth), where the identified PSMs match the ground truth sequence. Diagram **f** includes all query spectra, **g** shows spectra with multiple modifications, and **h~j** depict cases with single modifications at the N-terminal, middle, and C-terminal positions, respectively. **k** provides an example of a spectrum where our method performed better than ANN-SoLo, identifying a near C-terminus modification (formylation at T14) that ANN-SoLo missed. **l** displays a spectrum where b-ions are observed more frequently than in other spectra, with an AEBS modification located at K3. In this case, although the modification is near the peptide's N-terminus, the high abundance of b-ions causes nearly all b-ions to shift. This prevented ANN-SoLo from identifying the spectrum, whereas our method successfully identified it using the complementary database.

## 5.2 Analysis by Modification Position

We analyzed the identified PSMs based on the modification positions relative to the peptide sequence. The modifications were categorized into four groups:

1. N-terminal: Modifications near the start of the peptide.
2. Middle: Modifications located within the central region.
3. C-terminal: Modifications near the end of the peptide.
4. multiple modification: Peptide that has two or more modification

To understand the distribution of modification positions in the query spectra, we categorized all 225,336 spectra based on their modification locations and multiple modification:

1. N-terminal: 72749
2. Middle: 44314
3. C-terminal: 43375
4. multiple modification: 64898

We then compared the PSMs identified by ANN-SoLo and our method for each modification location. This analysis highlighted the relative strengths of each method in capturing modifications at different positions and identified the locations where our method exhibited the most significant improvement over ANN-SoLo. The specific differences are summarized in Figure 4. Since y-ions are considered dominant in spectra, modifications near the C-terminal, which cause a shift in almost all y-ions, cannot be accurately identified by ANN-SoLo. However, our method(fourth database) is specifically designed to account for such types of modifications(Figure 2B, Figure 4E,J), resulting in the most significant improvements in PSMs and correct sequence identifications at the C-terminal position.

It is possible that certain spectra exhibit more b-ion observations than y-ions. In such cases, the performance of our method depends on the modification position. If the modification is near the N-terminus, such spectra are likely to be identified in our complementary database(Figure 4L). Conversely, if the modification is located near the C-terminus, it may be identified using the original ANN-SoLo database. While our method is capable of identifying these spectra, ANN-SoLo may achieve higher identification rates due to its larger search space, as will be discussed in Section 5.4. Despite this, our method demonstrates robustness by

effectively accounting for various modification positions and their associated challenges.

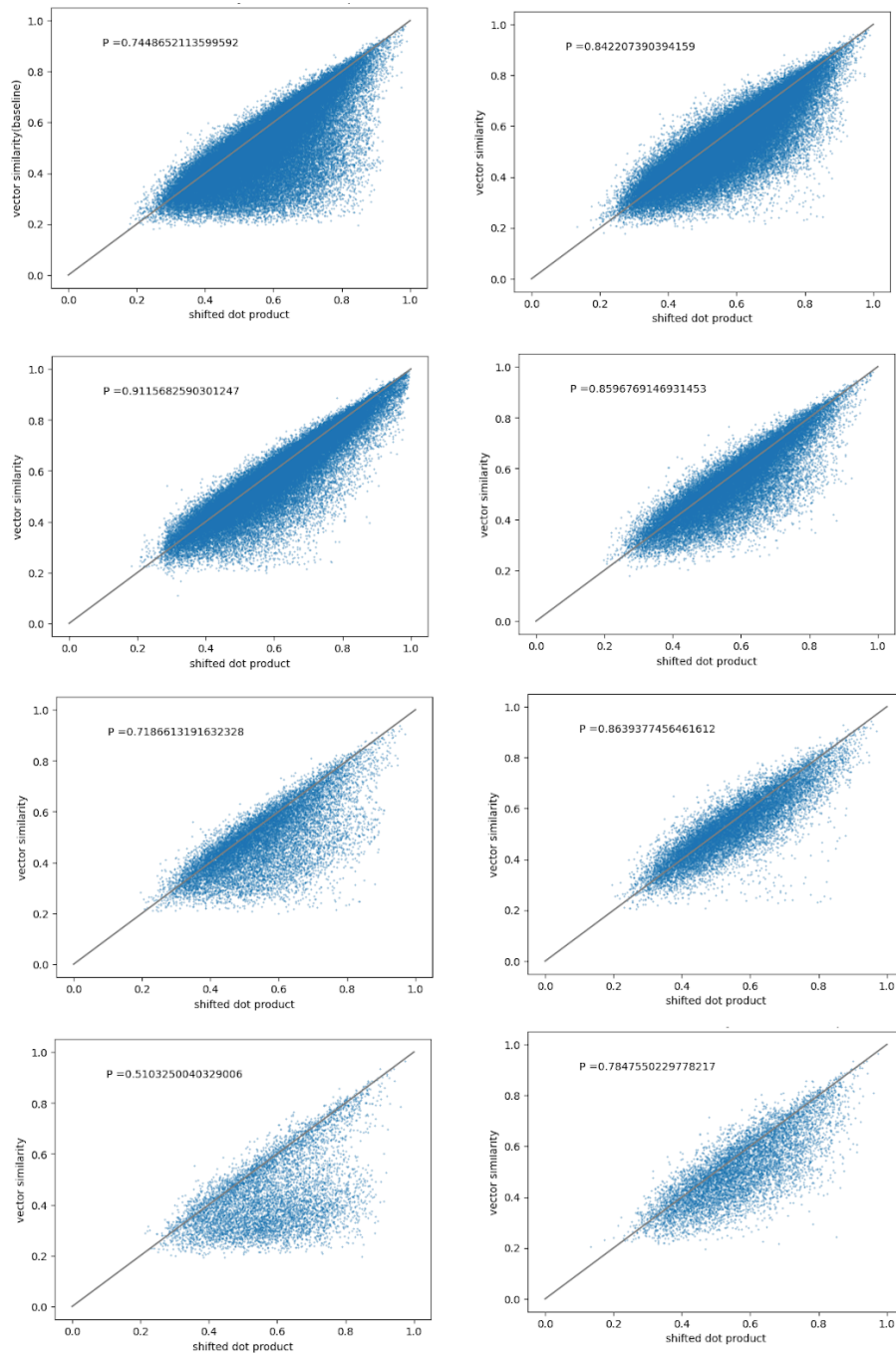




Figure 5. The left side shows scatter plots illustrating the correlation between the vector similarity (inner product, y-axis) obtained during candidate selection by ANN-SoLo and the shifted dot product (x-axis). The value PPP represents the Pearson correlation coefficient. From top to bottom, the plots represent the correlation for correct sequences identified through open search: the top plot shows all correct sequences, followed by cases where the modification is located near the N-terminus, in the middle, and near the C-terminus, respectively. The right side shows the corresponding scatter plots for correct sequences identified by our method. As expected, the correlation significantly differs for modifications located near the C-terminus, highlighting a key difference in performance between the two methods.

### **5.3 Correlation Analysis of Shifted Dot Product and Vector Similarity**

To further evaluate the efficacy of our method, we analyzed the correlation between the shifted dot product scores (used for candidate ranking) and the vector inner product scores (used for candidate selection). This analysis was conducted based on our belief that alignment between these two scores is critical to ensuring the validity of our approach.

As we expected, ANN-SoLo shows stronger correlation when modifications are located near the N-terminal and weaker correlation as modifications shift toward the C-terminal(Figure 5). However, in our method, the inner product results during candidate selection show a significant improvement in correlation when modifications are near the C-terminal. This observation supports the validity of our approach.

## 5.4 trade off between search space and search time

Increasing the Faiss parameter  $N_{probe}$  can yield more PSMs and higher accuracy. However, as shown in Figure 1, increasing  $N_{probe}$  results in a tradeoff where search time rises proportionally to the increase in  $N_{probe}$ . Our method demonstrated the ability to identify more PSMs and correct sequences compared to ANN-SoLo. Nevertheless, it is evident that some PSMs found exclusively by ANN-SoLo were not identified by our method.

This discrepancy arises from the lower  $N_{probe}$  setting used in ANN-SoLo to maintain reduced search time. By narrowing the search space, certain PSMs identified by ANN-SoLo were missed by our method. Despite this limitation, our method effectively identified PSMs across diverse modification positions, highlighting its capability to handle a variety of PTMs and modification positions. This demonstrates the robustness of our approach in addressing a wider range of modification scenarios.

## 5.5 PTM type

We analyzed the PTM types that our method identified more effectively compared to ANN-SoLo, as well as those it found difficult to detect. Among the PTM types that our method identified fewer of, deoxidation and Pyro-carbamidomethylation stood out. Based on the ground truth for correct sequences, ANN-SoLo identified 3,012 PSMs of deoxidation, while our method found 2,737. Similarly, for Pyro-carbamidomethylation, ANN-SoLo identified 2,290 PSMs, compared to 2,068 identified by our method.

This limitation in identifying certain PTM types can be attributed to our method's lack of significant improvement in performance when the PTM is located near the N-terminal, combined with ANN-SoLo's larger search space due to its parameter settings.

On the other hand, our method significantly outperformed ANN-SoLo in identifying formylation, with our method identifying 5,694 PSMs compared to ANN-SoLo's 5,241. In the case of AEBS modification, our method identified 6,165 correct PSMs, whereas ANN-SoLo identified 5,883. This result can be attributed to the fact that, in the query data, AEBS and formylation showed the largest difference in the number of peptides with modifications located near the C-terminal compared to those with modifications near the N-terminal.

Table1. PSM comparison on HEK293raw data

	overall	N-terminus	middle	C-terminus
ANN-SoLo	818,977	104,632	57,516	44,617
ours	827,715	104,924	57,990	45,339

Table 1. This table illustrates the performance differences between ANN-SoLo and our method. The results include both the standard search, which filtered out 534,576 PSMs, and the subsequent open search results. "Overall" represents all identified PSMs, while "N-terminus" refers to peptides with modifications near the N-terminus. "Middle" indicates modifications located in the middle of the peptide, and "C-terminal" refers to modifications near the C-terminus. As expected, the performance gap becomes more pronounced when the modification is located near the C-terminal.

Table2. correct sequence comparison on HEK293 raw data

	overall	N-terminus	middle	C-terminus
ANN-SoLo	550,473	82,741	50,292	38,977
ours	552,171	82,539	50,752	39,776

Table 2. This table presents the intersection of PSMs identified by ANN-SoLo and the ground truth sequences (considered as PSMs identified by MODa), as well as the intersection between our method and the ground truth sequences. "Overall" represents all identified PSMs, while "N-terminus" refers to peptides with modifications near the N-terminus. "Middle" indicates modifications located in the middle of the peptide, and "C-terminal" refers to modifications near the C-terminus.

## 5.6 Performance Comparison on HEK293 Raw Data

We compared the performance of ANN-SoLo and our method using a total of 1,121,149 spectra from HEK293, which included both unmodified and modified spectra. As a result, ANN-SoLo identified 818,977 PSMs, while our method identified 827,715 PSMs under a 1% FDR threshold. Additionally, when considering the PSMs identified by MODa as the ground truth (704,648 PSMs), ANN-SoLo identified 550,473 correct sequences, whereas our method identified 552,171 correct sequences. We have analyzed overall PSM performance and ground truth PSMs comparison considering modification position [Table 1,2].

## 6. Conclusion

ANN-SoLo is a tool used for open modification searches to identify post-translational modifications (PTMs). However, it had a limitation that the tool cannot consider modifications when selecting candidates. To address this, we developed a method that can account for modifications during candidate selection. This approach allowed us to find more peptide-spectrum matches (PSMs) without compromising execution time. Additionally, using ground truth data obtained through the tool MODa, we were able to identify more correct sequences.

We conducted an in-depth analysis based on the position of the modifications within the peptide sequences. Our method demonstrated significant improvements in identifying modifications near the C-terminus, a region where ANN-SoLo fails due to the dominance of y-ions in MS/MS spectra and the resulting shift of nearly all y-ions caused by such modifications. By effectively handling modifications in this region, our method made it possible to overcome this limitation of ANN-SoLo.

While our method shows improvements over ANN-SoLo, there is room for further refinement, particularly in handling cases with multiple modifications. Developing an approach that better considers and identifies spectra with multi-modifications could further enhance the algorithm's robustness and utility in future research.

## Reference

1. Ahrné, Erik, Markus Müller, and Frederique Lisacek. "Unrestricted identification of modified proteins using MS/MS." *Proteomics* 10.4 (2010): 671-686.
2. Ye, Ding, et al. "Open MS/MS spectral library search to identify unanticipated post-translational modifications and increase spectral identification rate." *Bioinformatics* 26.12 (2010): i399-i406.
3. Bittremieux, Wout, Kris Laukens, and William Stafford Noble. "Extremely fast and accurate open modification spectral library searching of high-resolution mass spectra using feature hashing and graphics processing units." *Journal of proteome research* 18.10 (2019): 3792-3799.
4. Dutta, Debojyoti, and Ting Chen. "Speeding up tandem mass spectrometry database search: metric embeddings and fast near neighbor search." *Bioinformatics* 23.5 (2007): 612-618.
5. Steen, Hanno, and Matthias Mann. "The ABC's (and XYZ's) of peptide sequencing." *Nature reviews Molecular cell biology* 5.9 (2004): 699-711.
6. Na, Seungjin, Nuno Bandeira, and Eunok Paek. "Fast multi-blind modification search through tandem mass spectrometry." *Molecular & Cellular Proteomics* 11.4 (2012).
7. Douze, Matthijs, et al. "The faiss library." *arXiv preprint arXiv:2401.08281* (2024).
8. Fu, Yan, and Xiaohong Qian. "Transferred subgroup false discovery rate for rare post-translational modifications detected by mass spectrometry." *Molecular & Cellular Proteomics* 13.5 (2014): 1359-1368.

9. Burke, Meghan C., et al. "The hybrid search: a mass spectral library search method for discovery of modifications in proteomics." *Journal of proteome research* 16.5 (2017): 1924-1935.
10. Wang, Mingxun, Jian Wang, Jeremy Carver, Benjamin S. Pullman, Seong Won Cha, and Nuno Bandeira. *Assembling the Community-Scale Discoverable Human Proteome*. Cell systems (2018).

## 국문요지

펩타이드 서열을 질량분석을 통해 해석하는 것은 단백질체학에서 필수적이면서도 복잡한 작업이다. 이를 위해 탠덤질량분석 데이터를 보다 효과적으로 해석하기 위한 다양한 알고리즘이 개발되어 왔다. 이 중 수식화 펩타이드의 동정을 위해 개방범위 탐색(Open Modification Search)을 활용하여 스펙트럼 라이브러리 탐색을 하는 방법이 최근 주목을 받고 있으며, 특히 근사 최근접 이웃 알고리즘(Approximate Nearest Neighbor)을 이용하여 후보 펩타이드를 군집화함으로써 기존 방법의 효율성을 높인 ANN-SoLo가 실용적인 결과를 제시하였다. 그러나 ANN-SoLo는 군집화 과정에서 펩타이드 수식화 정보를 고려하지 않아 스펙트럼의 해석과정에서 수식화된 펩타이드를 간과할 가능성이 존재한다. 이러한 한계를 극복하기 위해 본 연구에서는 보완 스펙트럼(complementary spectra)을 추가로 생성해서 군집화 과정에 수식화 정보를 반영하는 새로운 접근 방식을 제안하였다. 이를 통해 수식화 펩타이드와 수식화 되지 않은 펩타이드의 스펙트럼 쌍이 동일한 군집에 속하도록 하였다. 이처럼 수식화를 고려한 스펙트럼 클러스터링 방법은 수식화 스펙트럼의 동정을 위한 후보 펩타이드 선택 과정을 개선해, 후보 펩타이드 검색 속도를 보다 빠르게 했을 뿐 아니라, 스펙트럼 라이브러리 검색의 정확도를 향상시켜 단백질체학 연구에서 수식화 펩타이드 동정에 중요한 개선을 제공하였다.



## Declaration of Ethical Conduct in Research

I, as a graduate student of Hanyang University, hereby declare that I have abided by the following Code of Research Ethics while writing this dissertation thesis, during my degree program.

"First, I have strived to be honest in my conduct, to produce valid and reliable research conforming with the guidance of my thesis supervisor, and I affirm that my thesis contains honest, fair and reasonable conclusions based on my own careful research under the guidance of my thesis supervisor.

Second, I have not committed any acts that may discredit or damage the credibility of my research. These include, but are not limited to : falsification, distortion of research findings or plagiarism.

Third, I need to go through with Copykiller Program(Internet-based Plagiarism-prevention service) before submitting a thesis."

DECEMBER 01, 2024

Degree : Master  
Department : DEPARTMENT OF ARTIFICIAL INTELLIGENCE  
Thesis Supervisor : Paek, Eunok  
Name : SEO YOUNGHEE

(Signature)  


## 연구 윤리 서약서

본인은 한양대학교 대학원생으로서 이 학위논문 작성 과정에서 다음과 같이 연구 윤리의 기본 원칙을 준수하였음을 서약합니다.

첫째, 지도교수의 지도를 받아 정직하고 엄정한 연구를 수행하여 학위논문을 작성한다.

둘째, 논문 작성시 위조, 변조, 표절 등 학문적 진실성을 훼손하는 어떤 연구 부정행위도 하지 않는다.

셋째, 논문 작성시 논문유사도 검증시스템 "카피킬러"등을 거쳐야 한다.

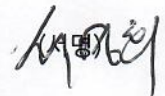
2024년12월01일

학위명 : 석사

학과 : 인공지능학과

지도교수 : 백은옥

성명 : 서영희



한 양 대 학 교 대 학 원 장 귀 하